Class-Based Parametric Approximation to Histogram Equalization for ASR

Luz García, Carmen Benítez Ortúzar, Angel De la Torre, and Jose C. Segura, *Senior Member, IEEE*

***Abstract*—This letter assesses an improved equalization transformation for robust speech recognition in noisy environments. The proposal is an evolution of the parametric approximation to Histogram Equalization named PEQ into a two-step algorithm dealing separately with environmental and acoustic mismatch. A first parametric equalization is done to eliminate environmental mismatch. These equalized data are divided into classes, and parametrically re-equalized using class specific references to reduce the acoustic mismatch. Experiments have been conducted for Aurora 2 and Aurora 4 databases. A comparative analysis of the experimental results shows significant benefits for databases with high acoustic variability like Aurora 4.**

***Index Terms*—Feature compensation, histogram equalization, parametric equalization, probabilistic classes, robust ASR.**

I. Introduction

WITHIN the group of feature normalization techniques for Robust Speech Recognition, statistical matching algorithms are very commonly used due to their low computational cost and the simplicity of their implementation. They eliminate the effects of noise in speech by modifying the statistics of the noisy feature vector, to make them equal to those of a set of clean reference vectors. Cepstral Mean and Variance Normalization (CMVN) [1] normalizes the two first statistical moments compensating the linear effects of additive noise in the Cepstral domain. A higher number of statistical moments have also been normalized [2], [3], achieving certain word error rate reductions with the drawback of a high computational cost. In that context, Histogram Equalization (HEQ) applied to voice features is the natural extension of the former efforts. It normalizes all the statistical moments of the Mel Frequency Cepstral Coefficients (MFCCs) by forcing their Cumulative Distribution Function (CDF) to match a clean reference CDF. In such way, the linear and nonlinear effects of noise which had modified the statistics and global shape of the histograms of the MFCCs are neutralized. Due to its simplicity (there is no assumption about the type or types of noises expected during recognition) and low computational cost, HEQ has been widely incorporated to the front-ends of speech recognizers in noisy environments [4]–[8]. Two inherent limitations must be pointed at in order to have its complete picture. Firstly, its "bag-of-frames" representation while equalizing a speech utterance implies the waste of the frame temporal context information. Secondly, it is not directly feasible for real-time processing because of the "look ahead" that would be required to calculate the utterance statistics.

The logic underneath HEQ is to transform train and test features to make them match a common range. Such *equalization* of ranges makes the features less vulnerable to acoustic and environmental mismatches, under the assumption of two premises. On one hand, the environmental mismatch acts as a monotonic invertible transformation in the feature domain. On the other, both train and test sentences contain enough acoustic information to provide acoustically accurate CDFs of the sentences to define properly the equalization transformation. The first premise is satisfied only partially. The random nature of noise adds a random, and by definition non invertible, transformation to the voice features that will not be eliminated by the equalization. There are some obstacles to fulfill the second assumption. For some ASR scenarios, the length of the sentences to be equalized is not enough to provide empirical representative acoustic statistics. Such lack of accuracy deteriorates the acoustic information of the equalized voice features which becomes then dependent on the particular content of the sentence. To overcome that limitation, parametric approximations like Double Gaussian Normalization (DGN) [9] or Parametric Equalization (PEQ) [10] have been proposed. They both assume two classes of speech frames (low energy and high energy frames) and use a parametric model of their probability distribution functions (pdfs). In that way, an equalization to a normal Gaussian reference CDF is performed, assuming also a Gaussian distribution for the observation features. Considering two classes of features and separately equalizing them facilitates the elimination of the acoustic mismatch between train and test sets. The usage of two separate independent histograms to equalize speech and silence frames was first proposed in [11] and later extended to a higher number of classes in the strategy named Probabilistic Class Equalization [12]. Such work uses the equalized version of the utterances to calculate class-probabilities for the corresponding noisy frames. Then, the original noisy sentences are equalized to multiple class-specific references. The final feature vector is calculated as the weighted average of all the class-specific equalizations.

This work progresses in the overcoming of the limitations of HEQ just defined, by extending the strategy of two-class Parametric Equalization (PEQ) proposed in [10] to a higher number of classes. Given the fact that acoustic classes are better identified once environmental mismatch has been eliminated, a

Manuscript received March 05, 2012; revised April 27, 2012; accepted May 05, 2012. Date of publication May 15, 2012; date of current version May 24, 2012. This work was supported by the Indo-Spanish Joint Programme of Cooperation in Science & Technology under Project ACI2009-0892 and by the Spanish Ministry of Science and Innovation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Murat Saraclar.

L. García, A. De la Torre, and J. C. Segura are with ETSIITT, Department of Signal Theory, Telematics and Communications, University of Granada, Granada, Spain 18071 (e-mail: mail: luzgm@ugr.es; atv@ugr.es; segura@ugr.es).

C. B. Ortúzar is with the Department of Electronics and Computer Technology, University of Granada, Granada, Spain 18071 (e-mail: carmen@ugr.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2012.2199485

first two-class parametric equalization is done. Then, these already equalized data are divided into a optimum higher number classes through unsupervised classification, and parametrically re-equalized using class-specific references. Two facts differentiate this proposal from [12]. Firstly, it deals with the class-specific equalization of features already cleaned through different normalization techniques. Secondly, parametric class-equalization is introduced. The number of parameters to be estimated is lower than in the case of the nonparametric approach, permitting to increase the number of classes used. As a result, the acoustic variability of the database is modelled with higher precision. The letter is organized as follows. Section II describes the procedure proposed. Section III presents the experimental work and numerical results obtained. Section III-A analyzes the benefits of the approximation for two databases with different acoustic variability: Aurora 2 and Aurora 4. In Section III-B, diverse global feature normalization techniques, including no normalization at all, are confronted to PEQ to evaluate the need of a first global noise reduction step and the appropriateness of PEQ as such. Finally, conclusions on the suitability of the approach are exposed in Section IV.

II. Class-Based Parametric Equalization

Speech features equalized to a reference distribution become invariant to arbitrary transformations as long as such transformations are invertible. Assuming that noise is an invertible transformation (ignoring the random part of its behavior) HEQ will make the equalized speech features invariant to noise.

Parametric models for the probability distributions of the features to be equalized have been proposed to reduce the statistical misestimations derived from the limited amount of observations per sentence in some scenarios. Empirical CDFs used in HEQ are substituted by Gaussian CDFs with a certain mean and variance in approximations that could also be considered as extensions of CMVN. Double Gaussian Normalization (DGN) uses a two-Gaussian mixture model (representing speech and nonspeech classes) for each component of the feature vector. A similar approximation named Parametric Equalization (PEQ) uses a multivariate probability model to separate speech and nonspeech frames instead of implementing an independent transformation for each component like DGN does. These techniques outperform successfully the existing HEQ mainly for two reasons. On one hand they provide an improvement in the statistical accuracy for the case of short test sentences with not enough frames to provide realistic cumulative histograms. On the other hand, they define independent transformations for speech and nonspeech frames making the equalization independent of the percentage of silence present in the sentence being equalized. This idea of class equalization is further extended to a higher number of acoustic classes in the work denominated Probabilistic Class Equalization. Such work presents satisfactory results by dividing the acoustic space into a set of acoustic classes and performing HEQ separately on each of them.

Based on the results just mentioned, the alternative proposed in this letter extends PEQ to a higher number of classes to analyze the convenience of implementing nonparametrically or parametrically the equalization. It could also be considered a class-based extension of CMVN on top of features already normalized.

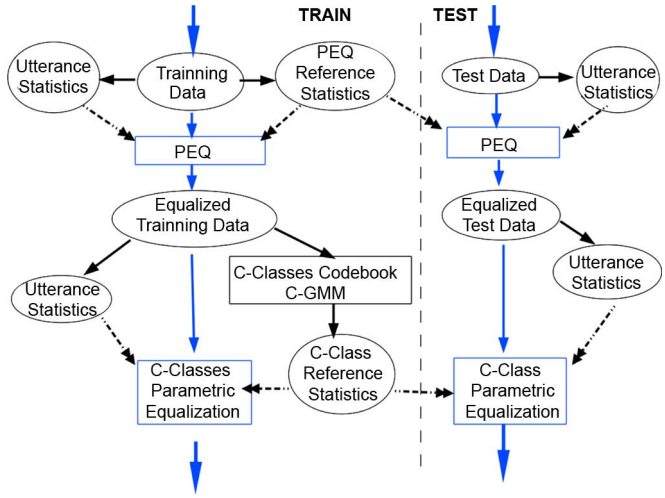

Fig. 1. Block diagram of the proposed equalization strategy.

Algorithm

Firstly, a two-classes Parametric Equalization (PEQ) is performed transforming a noisy feature vector y into a parametrically equalized feature vector $\hat{y}$. PEQ approximates the pdf of the MFCC features to a mixture of two Gaussians. The first Gaussian $N(\mu_n, \sigma_n)$ represents the pdf for nonspeech frames. The second Gaussian $N(\mu_s, \sigma_s)$ models the pdf for speech frames. Four steps are followed to obtain $\hat{y}$:

i) A two-Gaussian classifier based on the logarithmic energy of the frame (C0 Cepstral coefficient) is used to obtain the probability of each frame being a nonspeech frame $P(n|y)$, or a speech frame $P(s|y)$. These nonspeech and speech classes are initialized with frames having respectively their C0 Cepstral coefficient below (low energy) and above (high energy) the average C0 value of the sentence. Expectation and Maximization based re-estimations are done to obtain $P(n|y)$ and $P(s|y)$ values.

ii) A clean reference two-Gaussian pdf is calculated applying the classifier on the clean training data set to obtain the statistical references of the transformed domain $N(\mu_{nref}, \sigma_{nref})$ for the nonspeech class and $N(\mu_{sref}, \sigma_{sref})$ for the speech class.

iii) Using the Gaussian classifier based on C0, all frames of each feature vector y are classified as nonspeech or speech. The respective class-mean and class-variance for the utterance are calculated on vectors μ_{ny}, μ_{sy}, σ_{ny} and σ_{sy}.

iv) Each component of the feature vector y is linearly mapped to the clean reference statistic domain following (1) in case y is a nonspeech frame, or (2) in case it is a speech frame. The final parametrically equalized feature vector $\hat{y}$ is given by (3):

$$\hat{y}_n = \mu_{nref} + (y - \mu_{ny}) \cdot \left(\frac{\sigma_{nref}}{\sigma_{ny}} \right)^{1/2} \quad (1)$$

$$\hat{y}_s = \mu_{sref} + (y - \mu_{sy}) \cdot \left(\frac{\sigma_{sref}}{\sigma_{sy}} \right)^{1/2} \quad (2)$$

$$\hat{y} = P(n|y) \cdot \hat{y}_n + P(s|y) \cdot \hat{y}_s. \quad (3)$$

After this first global PEQ transformation, a second class-parametric equalization is done through the following steps:

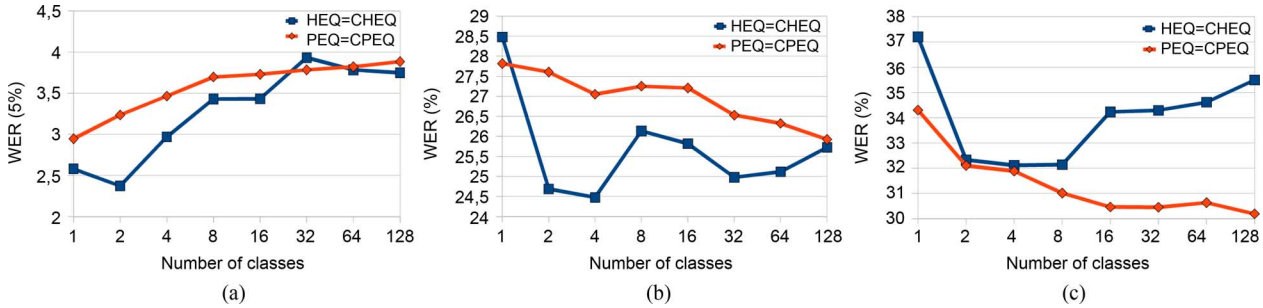

Fig. 2. Analysis for different databases and SNRs (a) Aurora 2: WER for Clean, 20 and 15 dBs (b) Aurora 2: WER for 10, 5 and 0 dBs (c) Aurora 4: WER.

v) A codebook of C centroids is calculated applying the k-means algorithm to the clean training dataset of vectors previously equalized with PEQ.

vi) Using such codebook and the clean training dataset of PEQ vectors, a *C-mixtures* Gaussian Mixture Model is estimated. Classes are initialized using the minimum euclidean distance to the C centroids. Expectation and Maximization based re-estimations are done to build the GMM model $N(\mu_{i_{ref}}, \sigma_{i_{ref}})$ for classes $i = 1$ to C.

vii) For every PEQ feature vector $\hat{y}$, mean $\mu_{\hat{y}}$ and variance $\sigma_{\hat{y}}$ are computed. The C-classes GMM created in step (vi) is first used to calculate the posterior probabilities of the vector $\hat{y}$ belonging to each of the classes $P(c_i|\hat{y})$, and secondly to linearly map $\hat{y}$ to each of the C classes following (4). The final feature vector is described in (5):

$$\hat{y}_{ci} = \mu_{i_{ref}} + (\hat{y} - \mu_{\hat{y}}) \cdot \left(\frac{\sigma_{i_{ref}}}{\sigma_{\hat{y}}} \right)^{1/2}, \quad i = 1 \; to \; C. \quad (4)$$

$$\hat{y}_c = \sum_{i=1}^{C} P(c_i|\hat{y}) \cdot \hat{y}_{ci}. \quad (5)$$

Fig. 1 shows a block diagram of steps (i)–(vii) followed through the train and test equalization process proposed.

III. Experimental Work

Class-based parametric equalization has been evaluated in the experimental framework of Aurora 2 [13] and Aurora 4 [14] databases following the standard clean-training tests. The procedures for training and recognition are identical to the reference experiments with the exception of the front-end that includes the feature normalization techniques described in this paper. Training and recognition are performed using the HMM Tool Kit (HTK) Software. A feature vector of 13 Cepstral coefficients is used as basic parametrization using coefficient C0 instead of the logarithmic energy. This basic feature vector is augmented with first and second order regressions yielding a final feature vector of 39 components. For comparison purposes, different parameterizations described in the following subsections have been implemented and evaluated.

A. Analysis for Different Databases

In order to evaluate the convenience of the parametric approximation proposed in this work, experiments have been conducted comparing it with an analogue implementation based on the traditional HEQ ([5], [12]). Two strategies have been confronted. Firstly, a quantile-based global histogram equalization followed by a class histogram equalization (named *HEQ-CHEQ*) has been performed. The dictionary of centroids and GMM used for the classification were built using the clean training set of features normalized with HEQ. Secondly, the two-class parametric equalization followed by the multi-class parametric equalization proposed in this letter has been implemented with name *PEQ-CPEQ*. Fig. 2 shows word error rate (WER) recognition results for experiments conducted on two databases with different acoustic variability and SNR conditions: Aurora 2 [13] and Aurora 4 [14]. Aurora 2 contains short utterances formed by connected digits artificially contaminated with several noises and SNRs ranging from clean condition to 0 dBs. Acoustic word models (16 states and 3 Gaussians per state) have been used for recognition. On the other hand, Aurora 4 is a large vocabulary database with bigger acoustic variability and several noises artificially added with an average SNR of 15 dB. Triphone models (3 states and 6 Gaussians per state) have been used for recognition.

Fig. 2 shows tendencies similar to those pointed at in the *Probabilistic Class Equalization* proposed for Aurora 2 in [12]. Fig. 2(a) plots the WER for Aurora 2 high SNR experiments (average results for clean, 20 and 15 dBs). The lowest WER is obtained using the nonparametric approach when two acoustic classes are separately equalized after a first global equalization *(HEQ-CHEQ)*. Plain HEQ and PEQ are also shown in experiment labeled ' *1 class*'. Fig. 2(b) depicts WER for low SNR Aurora 2 experiments (average results for 10, 5 and 0 dBs). Plain PEQ outperforms HEQ in noisy conditions but the optimal number of classes increases then to 4. The noisy acoustic environment is better modeled with a higher number of classes compared to the high SNR case. Aurora 4 is analyzed in Fig. 2(c) In this case, plain PEQ improves plain HEQ importantly (experiment "*1 class*"). Moreover, PEQ-CPEQ remarkably outperforms HEQ-CHEQ in all cases. Due to the raise in acoustic variability of Aurora 4, the optimal strategy is to significantly augment number of classes and use the class-based parametric equalization. When the number of classes increases, class dependent histograms become inexact. The less aggressive class-dependent mean and variance normalization done in the parametric approach produces better results then.

B. Optimum Domain for the Class Equalization

The objective of the PEQ transformation (performed as very first step of the algorithm proposed in Section II-A) is to reduce the environmental mismatch. Such noise removal improves itself recognition and permits calculating more exact statistics and probabilities for the acoustic classes of the utterances in the

TABLE I
AURORA 4 WER FOR DIFFERENT EQUALIZATION STRATEGIES

	Tests 01-07	Tests 08-14	Avrg.	R-I(%)
BASELINE	40.86	51.6	46.11	0
BASELINE-CHEQ	37.85	43.96	40.91	11.27
BASELINE-CPEQ	32.09	37.96	35.02	**24.05**
HEQ	32.28	41.87	37.08	19.58
HEQ-CHEQ	29.13	36.06	32.60	29.30
HEQ-CPEQ	29.17	35.66	32.41	**29.71**
PEQ	30.50	37.98	34.24	25.74
PEQ-CHEQ	29.65	33.86	31.75	31.14
PEQ-CPEQ	28.37	32.70	30.53	**33.79**
VTS	30.29	35.75	33.02	28.39
VTS-CHEQ	31.29	35.51	33.40	27.56
VTS-CPEQ	29.82	33.96	31.89	**30.84**

second equalization step. An analysis of the suitability of PEQ has been done based on experiments on Aurora 4 database.

Table I shows the WER obtained in recognition for several noise removal techniques applied to Aurora 4. Firstly, a baseline MFCC parameterization (*BASELINE* in the table) has been evaluated. Relative Improvement (R-I) over this baseline result is also shown in the table. Quantile based Histogram Equalization [5], named *HEQ*, has been implemented using a reference CDF averaged empirically over the whole clean training set. *PEQ* [10] has been performed computing both speech and nonspeech Gaussian probability density functions averaging the clean training data set. Finally, a third more complex feature transformation named Vector Taylor Series Approach *(VTS)* [15] has been applied. VTS uses an analytical expression of the environmental degradation based on the noisy observations and statistical models for the clean speech and the additive noise. Using such model of environmental mismatch it provides the expected value of the clean speech. A comparative analysis of these three feature normalization techniques shows that PEQ outperforms HEQ reducing its WER from 37.08% to 34.24% due to the introduction of the speech and nonspeech separate equalization and the parametric formulation. VTS outperforms both HEQ and PEQ lowering the WER to 33.02%.

WER results for Aurora 4 are also provided comparing the just mentioned normalization techniques as originary domains for the class equalization (baseline, HEQ, PEQ and VTS). Parametric (CPEQ) and nonparametric (CHEQ) class equalization of baseline features, and features transformed through HEQ, PEQ and VTS have been done. Results are shown for the number of classes empirically found to produce best average results using the database development set: eight classes in the case of baseline-CHEQ and baseline-CPEQ; four classes in the case of HEQ-CHEQ and HEQ-CPEQ; 32 classes in the case of PEQ-CHEQ and PEQ-CPEQ; 16 and four classes for VTS-HEQ and VTS-CPEQ respectively. The lowest WER is obtained when using CPEQ after PEQ, which is the topic of this letter. The PEQ originary domain outperforms the VTS originary domain, which seems interesting given the fact that plain VTS produces lower WER than plain PEQ. Performing a first statistical matching produces better global results when a second class-specific statistical matching is to be done on top of it. PEQ-CPEQ produces a R-I of 33.79% followed by the combination PEQ-CHEQ giving a R-I of 31.14%. The lowest R-I (11.27%) is obtained when applying directly class-equalization to baseline features. This result sustains the need of using some primary noise removal technique before doing the class-equalization.

IV. CONCLUSION

This letter proposes a class-based parametric approximation to the well known technique of Histogram Equalization for robust speech recognition. In order to improve the environmental and acoustic mismatch removal performed by HEQ, a two-steps strategy has been proposed. Firstly, a two-classes parametric equalization named PEQ is done to remove noise. Secondly, on such equalized domain acoustic classes are defined and features are normalized using class specific references. The introduction of acoustic classes separately equalized to match class specific statistics helps to overcome the acoustic mismatch in a very effective way. For databases with long sentences and high acoustic variability like Aurora 4 (large vocabularies), increasing the number of acoustic classes and simplifying the equalization to a parametric normalization of mean and variance produces satisfactory results that outperform the existing HEQ approximations.

REFERENCES

[1] O. Viiki, B. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," *Proc. ICASSP'98.*, 1998.
[2] I. M. Khademul and H. Keikchi, "On the effectiveness of MFCCs and their statistical distribution properties in speaker identification," in *Proc. IEEE Int. Conf. on Virtual Environments, Human Computer Interface and Measurement Systems*, 2004, pp. 136–141.
[3] H. Chang-wen and L. Lin-Shan, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition," *Proc. ICASSP'04*, 2004.
[4] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *IEEE Workshop on Automatic Speech Recognition and Understanding.*, 2001.
[5] J. C. Segura, C. Benítez, A. De la Torre, and A. Rubio, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 11, no. 5, pp. 517–520, 2004.
[6] A. De la Torre *et al.*, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, 2005.
[7] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 3, pp. 845–854, 2006.
[8] C. Wan and L. Lee, "Histogram-based quantization for robust and/or distributed speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 4, pp. 859–873, 2008.
[9] B. Liu, L.-R. Dai, J.-Y Li, and R.-H. Wang, "Double Gaussian based feature normalization for robust speech recognition," in *Proc. Int. Symp. on Chinese Spoken Language Processing.*, 2004.
[10] L. Garcia, J. C. Segura, J. Ramirez, A. De la Torre, and C. Benitez, "Parametric nonlinear feature equalization for robust speech recognition," *Proc. ICASSP'06.*, 2006.
[11] S. Molau, F. Hilger, D. Keysers, and F. Ney, "Enhanced histogram equalization in the acoustic feature space," *Proc. ICSLP'02.*, 2002.
[12] Y. Suh, J. Mikyong, and H. Kim, "Probabilistic class histogram equalization for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 4, 2007.
[13] D. Pearce and H. G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ICSLP'00.*, 2000.
[14] H. G. Hirsch, Experimental Framework for the Performance Evaluation of Speech Recognition Front-Ends of Large Vocabulary Tasks, STQ AURORA DSR Working Group, 2002.
[15] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, 1997, pp. 33–42.